心理学报 2023, Vol. 55, No. 10, 1712−1728
Acta Psychologica Sinica

https://doi.org/10.3724/SP.J.1041.2023.01712

模型参数点估计的可靠性：以 CDM 为例*

刘彦楼 [1,2] 陈启山 [3,4] 王一鸣 [2] 姜晓彤 [2]

([1] 曲阜师范大学教育大数据研究院; [2] 曲阜师范大学心理学院, 山东 济宁 273165)
([3] "儿童青少年阅读与发展"教育部哲学社会科学实验室(华南师范大学); [4] 华南师范大学心理学院, 广州 510631)

摘 要 心理学研究中, 不恰当的模型参数估计框架或收敛准则严重影响模型参数点估计的可靠性, 进而影响到研究结论的可靠性。本研究提出了基于 MLE-EM 的 CDM 模型参数估计新框架, 以及新收敛判断方法。通过模拟研究与实证数据分析的方式, 探索了新参数估计框架和新收敛判断方法的表现, 并与已有模型参数估计框架及收敛判断方法进行了比较。结果显示, 新的模型参数估计框架及收敛准则的表现优于已有的模型参数估计框架及收敛准则, 能有效提高模型参数点估计的可靠性。

关键词 参数估计, 点估计, 收敛准则, 认知诊断模型
分类号 B841

1 引言

自然科学及社会科学各个领域中, 研究结论的可靠性(研究结论可以被信赖的程度), 尤其是研究结果的可重复性(replication)受到极大关注(参见: 胡传鹏 等, 2016; Begley & Ellis, 2012; Ioannidis, 2005, 2008; Tajika et al., 2015)。Nature 杂志对此进行了一项调查, 发现 70%以上的研究者无法重复他人实验, 50%以上的研究者无法重复他们自己的实验(Baker, 2016)。心理学领域中, 研究者对可重复性问题出现的比例、可能的原因展开了探讨, 并从统计方法和研究实践两方面提出了解决方案(例如, 可参考《心理学报》的投稿指南及论文自检报告或 American Psychological Association, 2020 等)。

心理学研究中, 研究者使用模型参数描述被试的外显行为(或观察数据)与其潜在特质之间的关系。研究者在使用心理计量模型拟合数据时, 倾向于将计量模型及参数估计软件作为一个"黑箱"使用, 很少关注模型参数估计值是否可靠。举例而言, 极大似然法是当前应用最广泛的模型参数估计方法之一, 极大似然法中仅存在全局最优解的一个前提是似然函数是凸函数。然而, 实践中这个假设有可能不成立, 使得模型参数存在两个及以上的局部最优解。使用同一个模型分析相同数据时, 不同初始值可能会导致模型参数收敛于不同的局部最优解。根据极大似然法原理, 似然函数值不同, 说明产生了不同的模型参数估计值; 似然函数值之间的差异越大, 说明模型参数局部最优解之间的差异越大。例如, 假设 γ 是模型中任意一个参数, 如果第一次的点估计值与第二次的点估计值的差 $\hat{\gamma}^{(1)}-\hat{\gamma}^{(2)}$ 不近似为 0, 说明在这两次估计中模型参数 γ 的估计值及 95% CI 不同。

模型参数点估计的可靠性是研究结论可靠性的基础。因此, 如何提高模型参数估计值的可靠性, 进而提高研究结果的可重复性是本文将要探讨的主要问题。

认知诊断(或者是诊断分类)使用心理计量模型推断被试可观察的外显行为与其潜在的多维、细粒度的心理特质(如心理结构、技能、加工过程或策略等, 统称为属性)之间的关系(Rupp et al., 2010)。

收稿日期: 2023-03-02
* 国家自然科学基金青年项目(31900794)、山东省教育科学规划课题(2020KZD009)、广东省哲学社会科学规划项目(GD22CXL01)、广东省教育科学规划课题(2022GXJK176)和大学生创新创业训练计划(202110446231X)资助。
通信作者: 刘彦楼, E-mail: liuyanlou@163.com

认知诊断模型(cognitive diagnostic model, CDM)在心理、教育、社会、生物以及其他多个领域中得到了越来越多的关注(Sorrel et al., 2016; Wu et al., 2017)。因此，本文以 CDM 为例，探讨模型参数点估计的可靠性问题。

目前，极大似然期望最大化算法(maximum likelihood estimation using the expectation maximization algorithm, MLE-EM)是应用最广泛的 CDM 模型参数估计方法之一(de la Torre, 2009, 2011; von Davier, 2008)。例如，在 R 语言中的 *CDM* (George et al., 2016)、*GDINA* (Ma & de la Torre, 2020)软件包以及 *flexMIRT*, *Latent GOLD*, *mdltm*、*Mplus* (Sen & Terz, 2020; Templin & Hoffman, 2013)等软件中均可使用 MLE-EM 估计 CDM 的模型参数。理想条件下，使用 MLE-EM 方法能够获得具有渐近性、一致性等优良特性的点估计值。但是，研究者指出使用 MLE-EM 算法估计 CDM 模型参数时，可能会遇到的问题有：模型参数不收敛、项目参数极端值、(较差的)局部最优解以及边界值等(DeCarlo, 2011, 2019; Ma & Guo, 2019; Ma & Jiang, 2021; Philipp et al., 2018; Templin & Bradshaw, 2014; Zeng et al., 2023)。MLE-EM 估计的一般过程是，给定模型参数初始值，迭代进行 E 步(期望步)和 M 步(最大化步)，满足特定的收敛准则(convergence criterion 或 termination criterion)后停止迭代，输出模型参数的点估计值。因此，可以从参数估计框架(包括模型参数初始值设置、EM 过程等)及收敛准则等方面着手解决模型参数点估计可靠性问题。

本文将在第 2 部分阐述 CDM 模型参数估计中模型参数估计框架及收敛准则存在的问题，以及这两个问题对于参数估计可靠性的影响；在第 3 部分详细说明新提出的模型参数估计框架及收敛准则，并在第 4 部分通过模拟研究比较新方法与已有方法在模型参数估计可靠性方面的表现；第 5 部分是实证数据分析，目的是检验新提出的模型参数估计框架及收敛准则在估计 CDM 模型参数时的表现，并与 *GDINA* 软件包的表现进行比较；最后是讨论与展望。

2 CDM 及其模型参数估计中存在的问题

在这一部分，将首先介绍饱和 CDM 及属性层级 CDM(hierarchical cognitive diagnostic model, HCDM)；然后以此为基础阐述模型参数估计中存在的不收敛、项目参数极端值、(较差的)局部最优解以及边界值等问题。

2.1 饱和 CDM 及 HCDM

为表达便利，设在一个认知诊断测验中有 N 个被试，K 个属性，J 个项目，且属性与项目均为 0-1 计分。令矩阵 $\boldsymbol{y}=\{y_{nj}\}^{N\times J}$ 表示被试在测验项目上的观察作答反应，$y_{nj}=1$ 表示被试 n 正确作答项目 j，$y_{nj}=0$ 表示错误作答。矩阵 $\boldsymbol{Q}=\{q_{jk}\}^{J\times K}$ 表示属性与测验项目的对应关系，$q_{jk}=1$ 表示项目 j 测量了属性 k，$q_{jk}=0$ 则表示没有测量。矩阵 $\boldsymbol{\alpha}=\{\alpha_{lk}\}^{L\times K}$ 表示所有可能的属性掌握模式，$\alpha_{lk}=1$ 表示具有第 l 种属性掌握模式的被试掌握了属性 k，$\alpha_{lk}=0$ 表示没有掌握，L 表示所有可能的属性掌握模式的数量。参考以往研究的表述(田伟 等, 2014; Dempster et al., 1977)，将被试的项目反应矩阵 $\boldsymbol{y}$ 称为不完整数据(“incomplete” data)，将项目反应矩阵及被试的属性掌握模式组合而成的矩阵称为完整数据(“complete” data)。即完整数据矩阵 $\boldsymbol{x}$，可以表示为，

$$\boldsymbol{x}=\begin{pmatrix}\boldsymbol{x}_1\\ \vdots\\ \boldsymbol{x}_n\\ \vdots\\ \boldsymbol{x}_N\end{pmatrix}=\begin{pmatrix}\boldsymbol{y}_1 & \boldsymbol{\alpha}_1\\ \vdots & \vdots\\ \boldsymbol{y}_n & \boldsymbol{\alpha}_n\\ \vdots & \vdots\\ \boldsymbol{y}_N & \boldsymbol{\alpha}_N\end{pmatrix} \tag{1}$$

CDM 的结构模型定义了被试总体中所有可能的属性掌握模式 $\boldsymbol{\alpha}$ 的分布比例。令 $\boldsymbol{\pi}=(\pi_1,\cdots,\pi_l,\cdots,\pi_L)^{\mathrm{T}}$ 表示结构参数向量且 $\pi_L=1-\sum_{l=1}^{L-1}\pi_l$，$\pi_l$ 表示第 l 种属性掌握模式 $\boldsymbol{\alpha}_l=(\alpha_{1k},\cdots,\alpha_{lk},\cdots,\alpha_{lK})^{\mathrm{T}}$ 在被试总体中的分布比例，符号“T”表示转置。饱和 CDM 的结构模型中 $L=2^K$。CDM 的项目反应模型表示的是具有第 l 种属性掌握模式 $\boldsymbol{\alpha}_l$ 的被试 n 在测验项目 j 上的正确作答概率。饱和 CDM 项目正确作答的条件概率可以表示为，

$$P_{nj}=P(y_{nj}=1|\,\boldsymbol{\alpha}_l,\boldsymbol{q}_j)=\lambda_{j,0}+\sum_{k=1}^{K}\lambda_{j,1,(k)}\alpha_{lk}q_{jk}+\cdots\lambda_{j,K,(1,\cdots,K)}\prod_{k=1}^{K}\alpha_{lk}q_{jk} \tag{2}$$

公式(2)中 $\boldsymbol{q}_j=(q_{j1},\cdots,q_{jk},\cdots,q_{jK})^{\mathrm{T}}$ 表示项目 j 在 $\boldsymbol{Q}$ 矩阵中所对应的向量；在项目 j 中 $\lambda_{j,0}$ 表示截距项、$\lambda_{j,1,(k)}$ 表示对应于属性 k 的主效应项、$\lambda_{j,K,(1,\cdots,K)}$ 则表示最高阶交互效应项。项目参数向量 $\boldsymbol{\lambda}$ 及结构参数向量 $\boldsymbol{\pi}$ 构成了模型参数 $\boldsymbol{\gamma}=(\boldsymbol{\lambda}^{\mathrm{T}},\boldsymbol{\pi}^{\mathrm{T}})^{\mathrm{T}}$。

饱和 CDM 与 HCDM 的区别在于结构模型和项目反应模型的定义不同，HCDM 嵌套于饱和 CDM

中。为详细说明这两者之间的关系，现举例说明。假设一个测验中 $K=2, \boldsymbol{q}_j=(1,1)^{\mathrm{T}}$，被试 n 的属性掌握模式为 $\boldsymbol{\alpha}_n=(1,1)^{\mathrm{T}}$；且掌握第一个属性($\alpha_1$)是掌握第二个属性的前提($\alpha_2$)。那么，在饱和 CDM 中，所有可能的属性掌握模式可以表示为，

$$\boldsymbol{\alpha}=\begin{pmatrix}\boldsymbol{\alpha}_1^{\mathrm{T}}\\ \boldsymbol{\alpha}_2^{\mathrm{T}}\\ \boldsymbol{\alpha}_3^{\mathrm{T}}\\ \boldsymbol{\alpha}_4^{\mathrm{T}}\end{pmatrix}=\begin{pmatrix}0&0\\1&0\\0&1\\1&1\end{pmatrix} \tag{3}$$

即饱和 CDM 的结构参数可以表示为 $\boldsymbol{\pi}=\left(\pi_1,\pi_2,\pi_3,1-\sum_{l=1}^{3}\pi_l\right)^{\mathrm{T}}$；根据公式(2)，本例中的饱和 CDM 的项目反应函数可以表示为，

$$P_{nj}=\lambda_{j,0}+\lambda_{j,1,(1)}+\lambda_{j,1,(2)}+\lambda_{j,2,(1,2)} \tag{4}$$

根据属性层级关系，HCDM 中所有允许存在的属性掌握模式是，

$$\boldsymbol{\alpha}=\begin{pmatrix}\boldsymbol{\alpha}_1^{\mathrm{T}}\\ \boldsymbol{\alpha}_2^{\mathrm{T}}\\ \boldsymbol{\alpha}_4^{\mathrm{T}}\end{pmatrix}=\begin{pmatrix}0&0\\1&0\\1&1\end{pmatrix} \tag{5}$$

即 HCDM 的结构参数可以表示为 $\boldsymbol{\pi}=(\pi_1,\pi_2,1-\pi_1-\pi_2)^{\mathrm{T}}$；HCDM 中的项目反应函数可以表示为，

$$P_{nj}=\lambda_{j,0}+\lambda_{j,1,(1)}+\lambda_{j,2,(1,2)} \tag{6}$$

比较表达式(3)和(5)，及表达式(4)和(6)，可以发现将饱和 CDM 中的一些结构参数以及项目参数约束为 0，可获得 HCDM。也就是，如果“真”模型为 HCDM，但使用饱和 CDM 估计模型参数时，部分模型参数的真值等于 0。一些结构模型参数真值等于 0，意味着这些参数在参数空间的下界，如果不解决这种这类边界值问题可能会造成 MLE-EM 参数估计存在多种问题。

2.2 CDM 模型参数估计中可能存在的问题

使用 CDM 拟合作答反应数据时，如果模型参数过多、样本量较小，或者是模型参数中存在边界值尤其是结构参数中存在边界值等问题时，可能导致模型参数不收敛、项目参数存在极端值或者是存在多个局部最优解等问题(Ma & Jiang, 2021; Templin & Bradshaw, 2014)。

CDM 的项目正确作答概率及结构参数均介于[0,1]之间。在估计模型参数时可能会遇到项目参数或结构参数在参数空间的上界或下界的问题，这可能会造成模型参数无法估计，或者是造成模型参数的标准误过大甚至是无法求解。Ma 和 Jiang (2021)提出贝叶斯众数估计及单调约束，估计 G-DINA 模型的项目参数。但是，他们的研究指出贝叶斯众数估计或贝叶斯众数与单调约束结合的算法估计获得的项目参数可能是有偏的；另外，他们也指出在实践应用中先验分布的选择需要非常谨慎，因为不恰当的先验信息可能会导致误导性的、甚至是错误的结果。为将模型参数估计值约束在适当的边界中，Yamaguchi (2023)进一步提出将结构参数也要加以约束。然而，当属性之间存在层级关系，但是使用饱和结构模型估计参数时，有些结构参数的真值等于 0，以不恰当的先验约束使其远离 0 的做法是不对的。

使用 MLE-EM 估计 CDM 模型参数时需要设定模型参数初始值 $\boldsymbol{\gamma}^{(0)}$。在 $\boldsymbol{\gamma}^{(0)}$的基础上，E 步求完整数据似然函数的期望，M 步求最大化期望函数的模型参数。每一次迭代(记为，*rep*)中都会产生一个模型参数估计向量 $\boldsymbol{\gamma}^{(\mathrm{rep})}$，收敛判断方法的值小于收敛容差或者达到最大迭代次数则迭代停止(George et al., 2016; Ma & de la Torre, 2020)。如果是因为收敛判断方法的值小于收敛容差而停止迭代，那么模型参数收敛，并且将最后一次迭代中的参数作为模型参数估计值 $\hat{\boldsymbol{\gamma}}$；否则，没有收敛。

接下来将针对 CDM 模型参数估计中的结构参数边界值、最大迭代次数以及初始值可能对 M 步造成的影响及迭代次数展开探讨，阐述已有方法存在的问题。具体而言，MLE-EM 的 E 步中进行的是：给定观察数据 $\boldsymbol{y}$ 以及模型参数 $\boldsymbol{\gamma}^{(rep)}$ 条件下，求完整数据 $\boldsymbol{x}$ 对数似然函数的期望，

$$Q[\boldsymbol{\gamma}\mid\boldsymbol{y};\boldsymbol{\gamma}^{(rep)}]=\mathbb{E}_{\boldsymbol{x}\mid\boldsymbol{y},\boldsymbol{\gamma}^{(rep)}}\left\{\log\left[\prod_{n=1}^{N}\pi_n\prod_{j=1}^{J}P_{nj}^{y_{nj}}(1-P_{nj})^{1-y_{nj}}\right]\right\} \tag{7}$$

E 步除了获得以上表达式外，还根据观察数据 $\boldsymbol{y}$ 以及模型参数 $\boldsymbol{\gamma}^{(rep)}$ 计算出第 *rep* 次迭代中所有属性掌握模式的期望次数 $n_l^{(rep)}$ 以及每种属性掌握模式下正确作答项目 j 的期望人数 $r_{lj}^{(rep)}$。M 步进行的是：求最大化函数 $Q[\boldsymbol{\gamma}\mid\boldsymbol{y};\boldsymbol{\gamma}^{(rep)}]$ 的模型参数 $\boldsymbol{\gamma}^{(rep+1)}$。然后用 $\boldsymbol{\gamma}^{(rep+1)}$ 替换 E 步中的模型参数 $\boldsymbol{\gamma}^{(rep)}$，并依次迭代。直到满足收敛条件，或者达到了预先设定的最大迭代次数而停止。

以饱和 G-DINA 模型的参数估计为例，在 M 步中，经过公式推导(参考，de la Torre, 2009, 2011)可以求得更新后的第 l 种属性掌握模式下项目 j 正确作答概率的表达式，

$$P_{lj}^{(rep+1)} = \frac{r_{lj}^{(rep)}}{n_l^{(rep)}} \tag{8}$$

根据 $P_{lj}^{(rep+1)}$ 可以容易地获得 CDM 项目参数的估计值 $\boldsymbol{\lambda}_j^{(rep+1)}$；更新后的结构参数估计值可以表示为

$$\pi_l^{(rep+1)} = \frac{n_l^{(rep)}}{N} \tag{9}$$

CDM 研究中至少有两种情形的存在会使得结构参数出现边界问题(DeCarlo, 2011, 2019; Templin & Bradshaw, 2014; Yamaguchi, 2023)。第一种情形是属性之间存在层级关系，但使用饱和模型估计。对比饱和 CDM 及 HCDM 可以发现，如果“真”模型是 HCDM，但是用饱和 CDM 拟合数据的时候，模型中的一些结构参数是“不允许存在”的参数，即这些参数的真值为 0。第二种情形是样本量较少时可能使得某些属性掌握模式所对应的被试量较少或是等于 0。以上两种情形中，结构参数 π_l 的真值等于 0，由于 $n_l = N \times \pi_l$，可能使 M 步中出现属性掌握模式的期望数 $n_l^{(rep)}$ 等于 0 的问题。即公式(8)中分子、分母有可能等于 0，造成迭代异常终止。结构参数边界值问题与模型收敛判断以及 CDM 的参数估计的可靠性紧密关联。

对于边界值可能引起的问题，目前至少有 3 种解决方法。第一种是使用先验分布对正确作答概率加以约束(Liu et al., 2016; Ma & Jiang, 2021)。这种方法在使用时需要非常谨慎，因为它会导致有偏的参数估计值，尤其是在属性之间存在层级关系的情境中。第二种是 *GDINA* 软件包中默认采用的方法(Ma et al., 2022)。具体做法是：如果公式(8)的分母小于 0.001，那么在分子、分母上分别加校正系数 0.0005、0.001，即令 $P_{lj}^{(rep+1)} = 0.0005/0.001 = 0.5$。然而，这一设置是否合理有待商榷。第三种是 *CDM* 软件包中采用的方法，每次迭代中均在公式(8)的分母 $n_l^{(rep)}$ 加上一个非常小的值 10^{-10} (Robitzsch et al., 2022)。但是，这种设置在一些特殊情况下(如，分子、分母的值均接近 10^{-10} 时)是否合理同样有待商榷。

MLE-EM 在迭代进行前需要设置模型参数初始值。CDM 模型参数估计中参数初始值向量 $\boldsymbol{\gamma}^{(0)}$ 的设置可能会对 MLE-EM 的表现造成影响。估计模型参数时，MLE-EM 以参数初始值 $\boldsymbol{\gamma}^{(0)}$ 为起始点通过迭代逐渐收敛到(局部)最优的模型参数估计。理想情况下，函数表达式(7)中仅存在全局最优解，初始值 $\boldsymbol{\gamma}^{(0)}$ 不会对最终的模型参数估计值 $\hat{\boldsymbol{\gamma}}$ 产生影响。然而，当 CDM 的似然函数存在多个局部最优解时，初始值 $\boldsymbol{\gamma}^{(0)}$ 不同，最终估计获得的 $\hat{\boldsymbol{\gamma}}$ 也会不一样。即，当模型满足特定收敛准则时，模型参数估计值 $\hat{\boldsymbol{\gamma}}$ 可能仅是一个较差的局部最优解(Ma & Guo, 2019; Zeng et al., 2023)。为提高 CDM 模型参数估计值的可靠性，研究者提出使用多个初始值(例如，300)估计模型参数(Ma & Guo, 2019)；或者是生成多个初始值(例如，200)并计算其似然函数值，然后选择似然函数值最大的那组模型参数作为 MLE-EM 迭代的初始值。图 1 中呈现了单个参数的局部最优解与全局最优解的简单示例。在这个例子中，有两个点是局部最优解，一个全局最优解。假设 $\gamma^{(0)}$ 是 CDM 模型中的任意一个参数的初始值，如果 $\gamma^{(0)}$ 在 A 点，那么最终收敛于图中左侧的局部最优解；如果 $\gamma^{(0)}$ 在 C 点，那么最终收敛于图中右侧的局部最优解；如果 $\gamma^{(0)}$ 在 B 点，那么最终收敛于全局最优解；在这 3 个解中，初始值取 A 点时的解是最差的。需要特别说明的是，CDM 参数估计过程远比图 1 中呈现的过程复杂，单一的初始值难以保证获得较好的模型参数估计。

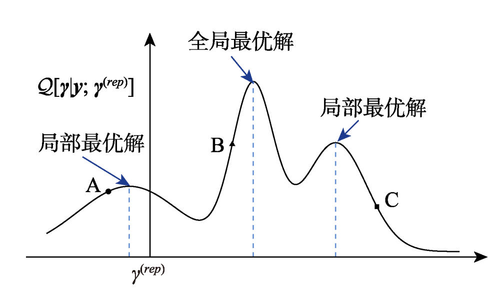


图 1 单个参数的局部最优解或全局最优解的简单示例

2.3 CDM 模型参数估计的收敛准则

收敛准则用于判断模型参数估计值是否已经足够接近模型参数最优解。一般而言，收敛准则由收敛判断方法、收敛容差及最大迭代次数这三部分组成(Paek & Cai, 2013)。收敛容差是研究者在模型参数估计前预先设定的、用于判断模型是否收敛的一个较小的值(例如，10^{-3} 或 10^{-6}，甚至更小)。模型参数估计中，如果实际迭代次数没有达到预先设定的最大迭代次数，收敛判断方法在迭代前与迭代后的差异小于收敛容差，说明模型参数估计值收敛；如果实际迭代次数达到了最大迭代次数，但是收敛判断方法在迭代前与迭代后的差异没有小于收敛容差，说明模型参数估计值没有收敛，无法获得模型的极大似然估计值。

当前，可以用于判断 CDM 模型参数估计是否

收敛的方法至少有 6 种(George et al., 2016; Ma & de la Torre, 2020; Ma et al., 2022; Robitzsch et al., 2022; Rupp & van Rijn, 2018)。

第一种是项目参数差的绝对值。这种方法的思想是如果迭代后的项目参数向量值 $\boldsymbol{\lambda}^{(rep+1)}$ 与迭代前的项目参数向量值 $\boldsymbol{\lambda}^{(rep)}$ 的差的绝对值中的最大值 $\max\{\text{abs}[\boldsymbol{\lambda}^{(rep+1)}-\boldsymbol{\lambda}^{(rep)}]\}$，小于预先设定的收敛容差，则认为模型参数收敛且停止迭代。这种收敛判断方法的优势在于，它所使用的收敛容差就是项目参数的精度。

第二种是模型参数差的绝对值。这种方法与项目参数绝对值差类似；不同之处在于模型参数差的绝对值将结构参数的差也纳入到收敛判断中，即 $\max\{\text{abs}[\boldsymbol{\gamma}^{(rep+1)}-\boldsymbol{\gamma}^{(rep)}]\}$ 中的最大值仍小于收敛容差时，认为模型参数收敛。

第三种是项目正确作答概率差的绝对值。这种方法比较的是迭代前后所有项目在所有属性掌握模式条件下的正确作答概率的绝对值的差中的最大值 $\max\{\text{abs}[\boldsymbol{P}^{(rep+1)}-\boldsymbol{P}^{(rep)}]\}$ 是否小于某个预先设定的收敛容差，其中 $\boldsymbol{P}\in\{P_{lj}\}^{L\times J}$ 。

第四种是项目正确作答概率和结构参数组成的向量的差的绝对值。这种方法以第三种方法为基础，将结构参数也纳入考虑，因此不再赘述。可以发现，以上 4 种收敛判断方法是基于全部或部分模型参数的。CDM 中项目正确作答概率一般是由项目参数组合而成，也就是说相对于项目参数而言，项目正确作答概率差这种方法更容易满足模型收敛准则。

第五种是对数似然函数差。对数似然函数差计算的是观察数据的第 *rep* 次及与第 *rep*+1 次迭代的负 2 倍对数似然函数的差的绝对值 $\text{abs}\{-2[\ell(\boldsymbol{\gamma}^{(rep+1)}|\boldsymbol{y})-\ell(\boldsymbol{\gamma}^{(rep)}|\boldsymbol{y})]\}$ 。这个方法认为迭代前后的对数似然函数值的差小于收敛容差时，似然函数取得了最大值。然而，有研究者指出这种方法的不足之处在于对数似然函数值的大小受到项目数量及被试量的影响，因此建议使用相对似然差。

第六种是相对似然差。相对似然差方法将对数似然函数的值也纳入到收敛判断方法的计算中。它试图消除对数似然函数值的大小对于收敛准则的影响。这种方法比较的是迭代前后两个似然函数的差与当前似然函数的比的绝对值是否小于预先设定的收敛容差。*GDINA* 软件包中使用的是 $\text{abs}\{2[\ell(\boldsymbol{\gamma}^{(rep+1)}|\boldsymbol{y})-\ell(\boldsymbol{\gamma}^{(rep)}|\boldsymbol{y})]/\ell(\boldsymbol{\gamma}^{(rep+1)}|\boldsymbol{y})\}$ (Ma et al., 2022)。这个方法的不足之处在于模型参数估计前 $\ell(\hat{\boldsymbol{\gamma}}\,|\,\boldsymbol{y})$ 是未知的，因此如何根据这个未知值而去预先设置恰当的收容差是这个方法存在的问题。

CDM 模型参数估计中，研究者使用的收敛判断方法、收敛容差及最大迭代次数上有明显差异。研究者经常使用的收敛判断方法是项目参数差的绝对值，且对应的收敛容差为 10^{-3} 或 10^{-4} (参考, de la Torre 2009, 2011; Ma & de la Torre, 2016; Paulsen & Valdivia, 2022; Sen & Terzi, 2020)。一些研究者在使用项目参数差的绝对值时，将收敛容差设置的更小，例如 10^{-5} (George et al., 2016)、10^{-6} (Rupp & van Rijn, 2018)或 10^{-7} (Chiu et al., 2023)；也有一些研究者使用对数似然函数差进行收敛判断，并将收敛容差设置为 10^{-2} 或 10^{-3} (Khorramdel et al., 2019; Ma & Guo, 2019)。但是 Rupp 和 van Rijn (2018)认为对数似然函数差依赖于项目数量及被试量，在进行模型参数收敛判断时相对似然差可能会更好。但是他们并没有对相对似然差的表现，以及这种方法适用的收敛容差进行研究。

另外，研究者在估计模型参数时大多倾向于使用软件的默认设置，较少对默认选项进行修改，但是 CDM 模型参数估计软件的默认设置也有较大区别。举例而言，*GDINA* 及 *CDM* 软件包中默认使用的收敛准则有明显区别(Ma et al., 2022; Robitzsch et al., 2022)。*GDINA* 软件包中默认使用的收敛判断方法、收敛容差及最大迭代次数分别是：项目正确作答概率和结构参数组成的向量的差的绝对值、10^{-4} 及 2000。*CDM* 软件包中使用的是收敛方法的组合，并且不同函数使用的默认设置不同。*CDM* 软件包中 gdina 函数中默认使用的收敛准则是：收敛容差为 10^{-4} 的项目参数差的绝对值方法与收敛容差为 10^{-1} 的对数似然函数差方法的组合，且最大迭代次数为 1000。

可以发现，研究者使用的收敛准则有很大差别。因此，相同计量模型条件下，不同的收敛准则是否会对模型参数点估计的可靠性产生影响；如果产生影响，在目前所有可用的模型参数估计收敛判断方法中，哪种效果是最好的；或者是能否开发一种具有广泛适用性的方法提高 CDM 模型参数点估计的可靠性是一个需要解决的重要问题。

3 新的模型参数估计框架及收敛准则

如前所述 CDM 模型参数估计中的边界值、局部最优解、项目参数极端值、模型参数不收敛，以及收敛准则设置等可能会对模型参数点估计的可

靠性产生影响，进而可能会影响到研究结果的可重复性。因此，本文提出新的模型参数估计框架试图解决 2.2 部分提及的模型参数估计中可能存在的问题；提出新的收敛准则试图解决 2.3 部分提及的收敛准则可能存在的问题。

首先，阐述边界值问题的解决方法。通过 2.2 部分可以发现，当前关于边界值的 3 种解决方法都存在一些可能的不足。借鉴 *GDINA* 及 *CDM* 软件包中的设置，本文使用的是：如果公式(8)的分母小于 10^{-16} 时，在分母上加上 10^{-14}。第 l 种属性掌握模式下，正确作答项目 j 的期望人数(分子)不大于这个属性掌握模式下的期望人数(分母)，所以使用这个方法可以保证公式(8)中 $P_{lj}^{(rep+1)}$ 的最大值不会超过 0.01。即，这个方法在保证分母不等于 0 的前提下，尽量减小校正系数对正确作答概率的影响。感兴趣的读者可以尝试使用其他值，但是我们认为只要满足分母不等于 0，且 $P_{lj}^{(rep+1)}$ 较小(如，小于 0.01)这两个条件，不同的校正系数对模型参数估计结果不会产生明显影响。

其次，阐述局部最优解、项目参数极端值、模型参数不收敛等问题的综合解决方法。

模型参数收敛判断中，设置最大迭代次数的唯一目的是避免模型参数估计程序陷入到无限(或近乎于无限)循环。然而，在模型参数本应收敛的情况下，如果将最大收敛次数设置的过小，可能会使得 MLE-EM 过早结束循环，造成不收敛的错误结果。解决不收敛问题的首要一步是设置足够大的收敛次数，因此本研究中将最大收敛次数设置为 50000。

CDM 的模型参数仅存在全局最优解的一个前提是公式(7)为凸函数。但是，这个前提有时未必成立，导致模型参数可靠性变差。因此，参考 Ma 和 Guo (2019)的相关研究，本文提出使用多个初始值计算 CDM 模型参数。即，遇到不收敛或项目参数存在极端值时重新生成初始值并计算，如果新初始值条件下的模型参数收敛、对数似然函数值大于先前的值、且项目参数不存在极端值时，使用新的估计值作为最终的模型参数估计值。在接下来的部分将这个新的模型参数估计框架称为 *mCDM*，并以此为基础探讨各种收敛准则的表现。由于 *mCDM* 在特定条件下，需要对于同一观察数据矩阵 $\boldsymbol{y}$，在多个不同初始值下进行模型参数估计，运算量可能会比较大。因此参考以往研究(刘彦楼，2022)，*mCDM* 程序计算量大的部分采用 C++语言及并行计算进行。特别说明的是，*mCDM* 程序已上传到科学数据银行，感兴趣的读者可以自行下载使用。

最后，阐述本文中新提出的收敛判断方法。

极大似然法估计的原理是找到最大化观察数据对数似然函数的模型参数值，并将其作为模型参数“真值”的估计。收敛判断方法的用途是判断观察数据对数似然函数的值是否已经近似达到了最大。但是，单一的判断方法在特定条件下可能存在缺陷。以对数似然函数差及模型参数差的绝对值为例进行说明。对数似然函数差方法假定第 *rep* 次及与第 *rep*+1 次迭代的对数似然函数的差小于预设的收敛容差时，似然函数值达到了最大。图 2 中呈现了对数似然函数差收敛判断方法可能存在的缺陷的简单示例。假定 B 点为 CDM 中任意一个参数的初始值 $\gamma^{(0)}$。当模型参数 $\gamma^{(rep)}$ 接近全局最优解时，如果似然函数的曲线比较平坦(可参考 Farrell & Lewandowsky, 2018)，那么将会出现模型参数差的绝对值变化较大，但是对数似然函数差变化非常小的问题。即，模型参数差的绝对值的判断效果优于对数似然函数差。模型参数差的绝对值可能存在的问题在于，似然函数值的大小除了受到模型参数值的影响之外，还受到项目数量及被试数量的影响(可参考 Rupp & van Rijn, 2018)。

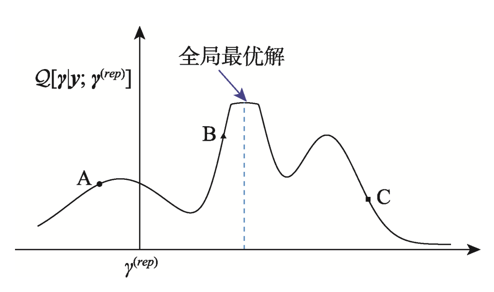


图 2　对数似然函数差收敛判断方法可能缺陷的简单示例

理论而言，进行 CDM 模型参数估计时，模型参数估计收敛判断方法及收敛容差设置越严格(这也就意味着在相同收敛容差条件下，迭代次数更多)，就越能获得使得 $\ell(\hat{\gamma}|\mathbf{y})$ 最大化的模型参数估计值。然而，实践中由于样本量、项目数量、属性数量、项目反应模型、属性层级关系及 $\boldsymbol{Q}$ 矩阵元素可能存在错误设定等因素的存在，很难预先判断哪种方法及相应的收敛容差是最严格的。因此，参考以往研究(George et al., 2016; Ma & de la Torre, 2020; Ma et al., 2022; Robitzsch et al., 2022; Rupp & van Rijn, 2018; von Davier, 2008; Xu & von Davier, 2008)，为克服单一判断方法可能存在的缺陷，本文提出在给定收敛容差的基础上综合使用模型参

数差的绝对值、项目正确作答概率和结构参数组成的向量的差的绝对值、对数似然函数差以及相对似然差进行模型参数收敛判断，并将其称为综合判断法。

需要说明的是，相对于项目参数(或项目正确作答概率)而言，结构参数的数量相对较少。被试观察作答反应数据可以为每个结构参数提供更多的信息，其估计值能够较快地固定下来。即，理论上收敛判断方法中是否包含结构参数应该没有明显差别。但是审慎起见，在接下来的研究中采用包含结构参数的方法。另外，与 *GDINA* 包不同，本研究中 *mCDM* 程序使用的相对似然差的计算公式是 $\text{abs}\{[\ell(\boldsymbol{\gamma}^{(rep+1)}|\boldsymbol{y})-\ell(\boldsymbol{\gamma}^{(rep)}|\boldsymbol{y})]/\ell(\boldsymbol{\gamma}^{(rep+1)}|\boldsymbol{y})\}$。

综上所述，本研究提出了基于 MLE-EM 的 CDM 模型参数估计新框架及新收敛准则，以提高模型参数点估计的可靠性。新的模型参数估计框架包括对 MLE-EM 方法中的 E 步及 M 步的改进。对 E 步的主要改进是，必要时(如，模型参数不收敛或项目参数存在极端值时)使用不同的初始值分别重新计算 E 步中的期望次数及进行后续的迭代。对 M 步的主要改进是，保证公式(8)中分母不等于 0 且 $P_{lj}^{(rep+1)}$ 较小。

4 模拟研究

4.1 研究目的

本研究重点关注的问题是：新提出的模型参数估计框架及收敛准则能否有效提高模型参数点估计值的可靠性。即，新提出的 *mCDM* 框架下的综合判断方法是否优于现有框架下的方法，能否在尽量保证参数在合理范围内的前提下，获得使得似然函数最大的参数估计值。具体包括：(1)数据生成模型与拟合模型均为饱和 G-DINA 时，即模型完全正确设定条件下各种收敛准则的表现；(2)数据生成模型为 HCDM 但使用饱和 G-DINA 拟合时，即模型中存在边界值时各收敛准则的表现。

4.2 研究方法

模型参数收敛准则的表现依赖于具体的模型参数估计方法，除本文中新开发的 *mCDM* 程序外，*CDM* (version 8.2-6; Robitzsch et al., 2022)、*GDINA* (version 2.9.3; Ma et al., 2022)这两个开源软件包也可用于模型参数估计。然而，*CDM* 包默认的设置是，当 $K \geq 4$ 时使用属性掌握模式简化方法估计结构参数(Xu & von Davier, 2008)。本文预研究发现这个方法下获得的一些结构参数估计值是有偏的。因此，本研究使用的模型参数估计框架有两种：*GDINA* 及 *mCDM* 程序。收敛判断方法有 5 种：模型参数差的绝对值、项目正确作答概率和结构参数组成的向量的差的绝对值、对数似然函数差、相对似然差以及综合判断法。并将这 5 种收敛判断法分别简记为：dp、ip、ll、rl 及 comp。参考先前研究，本文中考虑了 3 种收敛容差：10^{-4}、10^{-6}、10^{-8}。为区分不同的收敛准则，将 *GDINA* 框架下的收敛方法简称前加字母“G”、*mCDM* 框架下的方法加字母“m”，并将 3 种收敛容差的小数位数加在收敛方法简称后。例如，将 *GDINA* 框架下判断方法为模型参数差的绝对值及收敛容差为10^{-4}的收敛准则，简记为 Gdp4；将 *mCDM* 框架下判断方法为综合判断方法及收敛容差为10^{-6}的收敛准则，简记为 mcomp6。即，本文探讨 2 种计算框架、5 种收敛判断法、3 种收敛容差所组成的 30 种收敛准则在可能影响因素中的表现。

模拟研究中考虑了 2 种数据生成模型：饱和 G-DINA 模型以及属性(α_1、α_2、α_3)之间呈线性层级关系的 HCDM。鉴于实践中难以在 CDM 的模型参数之前预先设定恰当的层级关系，因此，选择饱和 G-DINA 作为拟合模型。样本量及项目数量对模型参数估计准确性有重要影响，因此对于收敛准则的表现也可能会产生影响。本研究中考虑了 3 种样本量：N = 500、1000 及 4000；有 2 个水平的项目数量：J = 16、32；且将属性数量固定为 4。为保证 CDM 的模型参数具有可识别性(Gu & Xu 2019, 2020)，项目数量为 16 时本研究使用图 3 中呈现的 $\boldsymbol{Q}$ 矩阵；将图 3 中的 $\boldsymbol{Q}$ 矩阵重复两次，构建项目数量为 32 时的 $\boldsymbol{Q}$ 矩阵。

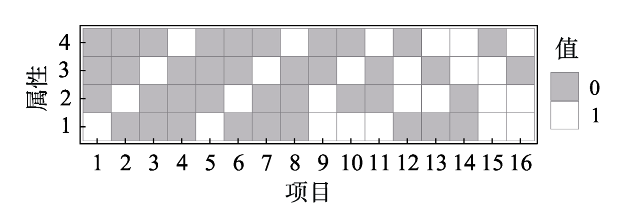


图 3 模拟研究中 J = 16 的 $\boldsymbol{Q}$ 矩阵

为更好地贴近 CDM 应用情景，参考 Liu (2018)及 Liu 等人(2022)的研究设计，使用以下步骤生成项目参数及结构参数真值：(1)项目参数中截距项(即猜测参数) $P(0)$ 随机取自[0.05, 0.4]的均匀分布；正确作答概率参数 $P(1)$ 随机取自[0.6, 0.95]；并且将主效应项及交互效应项设置为相等，即主效应及交互效应的参数值都等于$[P(1)-P(0)]$除以它们的个数。(2)结构参数的真值根据多维正态分布生成，

具体步骤是：首先，将多维正态分布的均值向量设置为 **0**，方差—协方差矩阵的非对角线元素的值从均匀分布[0.3，0.7]中随机抽取。然后，从多维正态分布中随机一百万被试，并以 0 为切点对每个被试的取值向量进行二分化处理，即，向量中的值大于 0 设置为 1，其他情况设置为 0，以此转化为属性掌握模式。最后，当数据生成模型为饱和 G-DINA 时，直接计算这一百万个被试的属性掌握模式在 L 种属性掌握模式上的分布比例，并将其作为结构参数的真值；当数据生成模型为HCDM时，只计算这一百万个被试的属性掌握模式中允许存在的属性掌握模式的比例，并将其作为 HCDM 中结构参数的真值。每种实验条件组合重复 500 次以获得稳定的模拟结果，且将 *mCDM* 及 *GDINA* 的最大迭代次数都设置为 50000。

4.3　评价指标

收敛准则的目的是判断迭代过程中的模型参数是否已经最大化了似然函数，因此本研究的评价指标主要围绕对数似然函数进行构建，包括：最佳似然函数次数(LL_{Best})，似然函数的均值(LL_{mean})、似然函数的最大值(LL_{max})、似然函数的最小值(LL_{min})以及似然函数的标准差(LL_{sd})。最佳似然函数次数指的是 30 种收敛准则在 500 次重复中分别取得最佳似然函数值的次数 $LL_{Best}=\sum_{R=1}^{500} I(LL_{Conv_R}=LL_{max_R})$；其中，$LL_{Conv_R}$ 表示各收敛准则在第 R 次重复中对应的对数似然函数值，$LL_{max_R}=\max(LL_{Conv_R})$ 表示第 R 次重复中所有收敛准则对应的对数似然函数的最大值，I 是示性函数用于判断前后两个函数值是否相等，如果 LL_{Conv_R} 与 LL_{max_R} 相等，函数 I 的值等于 1，否则等于 0。关于 LL_{Best} 需要特别说明的是，在单次循环中可能会有多个收敛准则同时取得最佳似然函数值；LL_{Best} 的值越大说明的是收敛准则的表现越好。LL_{mean}、LL_{max}、LL_{min} 以及 LL_{sd} 表示 500 次重复中 30 种收敛准则分别对应的对数似然函数值的均值、最大值、最小值以及标准差，例如 $LL_{mean}=mean(LL_{Conv_R})$。

其他评价指标还包括：500 次重复中 30 种收敛准则分别对应的模型参数估计程序单次运行的平均时间(t_{mean}，单位是秒)，平均迭代次数(Itr_{mean})，实际迭代次数的最大值(Itr_{max})，所有项目参数出现极端值的总数(将项目参数大于 1 或者是小于−1 定义为极端值，表示为 λ_{out})，以及模型参数估计程序未收敛次数的总次数。

4.4　模拟结果

在呈现具体结果前，首先对两个一般性的结果进行说明。本研究中所有实验条件组合下模拟结果显示：(1)当最大迭代次数为 50000 时，所有重复中的模型参数都收敛了，没有出现未收敛情况。即，未收敛次数指标均为 0。(2)在相同的模型参数估计框架(*GDINA* 或 *mCDM*)及收敛容差(10^{-4}、10^{-6}或 10^{-8})下，相对于其他方法而言，rl 方法的表现是最差的，其 LL_{Best} 均为 0。因此，在结果部分不再呈现 rl 方法的模拟结果。

4.4.1　饱和 CDM 生成数据时各收敛准则的表现

表 1 中呈现的是使用饱和 G-DINA 生成数据，$J=16$，$N=500$ 条件下除了 rl 方法外的 24 种收敛准则的表现。通过表 1 中的 LL_{Best} 指标可以发现，在这些收敛准则中，表现最好的是新框架 *mCDM* 下收敛容差为10^{-8}的综合判断法 mcomp8。就收敛判断方法而言，在相同收敛容差条件下，不论是 *GDINA* 框架还是 *mCDM* 框架下，表现最好的是 comp 方法，其次是 dp 方法；ip 与 dp 方法的表现较为类似，但是 dp 的表现稍好，这主要是因为 dp 是模型参数，而 ip 是参数的组合。就收敛容差而言，同一模型参数估计框架和收敛判断方法下，随着收敛容差变小，收敛准则的表现也在变好。以 comp 方法为例，随着收敛容差从10^{-4}变化到10^{-6}，Gcomp 在 LL_{sd} 指标上近似相等但是在 LL_{Best}、LL_{mean}、LL_{max}、LL_{min} 指标上的表现在变好；mcomp 在这些指标上的表现与 Gcomp 类似。另外需要指出的是，当收敛容差从10^{-6}变化到10^{-8}时，LL_{Best}、LL_{mean}、LL_{max}、LL_{min} 等指标几乎没有明显变化，但是可以发现 Itr_{mean} 以及 Itr_{max} 有较大增长。就模型参数估计框架而言，可以明显发现各个收敛准则在 *mCDM* 框架下的表现要优于 *GDINA* 框架，一个明显的例子是，mcomp8 在 LL_{Best} 指标上的表现优于 Gcomp8。并且 *mCDM* 框架下 t_{mean} 及 λ_{out} 也明显优于 *GDINA* 框架。通过 Itr_{max} 指标可以发现，无论是 *GDINA* 还是 *mCDM* 的实际所使用的迭代次数的最大值都超过了 30000。这说明一些 CDM 参数估计软件中最大迭代次数默认设置是不合理的，会产生模型参数不收敛的错误结论。

根据表 1 中的结果，各收敛判断方法的收敛容差等于10^{-4}时，在 LL_{Best} 指标上均没有好的表现；尽管 ip 类方法与 dp 类方法表现类似，但是 ip 类的表现相对较差。因此，模型完全正确设定条件下，不再呈现收敛容差为10^{-4}、及收敛判断方法为 ip 时，各收敛准则的模拟结果。

表 1 饱和 CDM 生成数据, $J = 16, N = 500$ 条件下的模拟结果

收敛准则	LL_{Best}	LL_{mean}	LL_{max}	LL_{min}	LL_{sd}	t_{mean}	Itr_{mean}	Itr_{max}	λ_{out}
Gdp4	0	−4948.024	−4847.235	−5054.561	34.436	0.540	180	848	62
Gdp6	240	−4948.011	−4847.226	−5054.557	34.437	1.181	474	5752	61
Gdp8	280	−4948.011	−4847.226	−5054.557	34.437	2.068	901	32057	61
Gip4	0	−4948.027	−4847.234	−5054.561	34.438	0.507	164	730	59
Gip6	232	−4948.011	−4847.226	−5054.557	34.437	1.131	452	5680	61
Gip8	279	−4948.011	−4847.226	−5054.557	34.437	1.847	863	28030	61
Gll4	0	−4948.024	−4847.229	−5054.558	34.438	0.520	169	844	60
Gll6	48	−4948.017	−4847.226	−5054.557	34.431	0.858	329	1819	61
Gll8	273	−4948.011	−4847.226	−5054.557	34.437	1.217	531	6760	61
Gcomp4	0	−4948.022	−4847.229	−5054.558	34.436	0.566	190	848	62
Gcomp6	240	−4948.011	−4847.226	−5054.557	34.437	1.189	478	5752	61
Gcomp8	281	−4948.011	−4847.226	−5054.557	34.437	2.062	905	32057	61
mdp4	0	−4948.021	−4847.234	−5054.560	34.436	0.254	179	877	59
mdp6	360	−4948.008	−4847.226	−5054.556	34.437	0.461	479	5803	59
mdp8	498	−4948.008	−4847.226	−5054.556	34.437	0.735	953	32053	59
mip4	0	−4948.022	−4847.234	−5054.560	34.436	0.241	165	774	58
mip6	346	−4948.012	−4847.226	−5054.556	34.441	0.432	453	5730	59
mip8	496	−4948.008	−4847.226	−5054.556	34.437	0.690	912	28026	59
mll4	0	−4948.021	−4847.228	−5054.557	34.437	0.240	168	923	57
mll6	69	−4948.018	−4847.226	−5054.556	34.435	0.349	335	1978	59
mll8	485	−4948.008	−4847.226	−5054.556	34.437	0.495	585	6756	59
mcomp4	0	−4948.019	−4847.228	−5054.557	34.435	0.258	189	923	59
mcomp6	363	−4948.008	−4847.226	−5054.556	34.437	0.462	485	5803	59
mcomp8	500	−4948.008	−4847.226	−5054.556	34.437	0.734	958	32053	59

表 2 中呈现的是使用饱和 CDM 生成数据, $J = 16$ 时 $N = 1000$ 和 4000 两种样本量水平下各收敛准则的表现。在 $N = 1000$ 样本量水平下, 表现最好的收敛准则同样是 mcomp8; 当 $N = 4000$ 且收敛容差为 10^{-8} 时, 表 2 中各收敛准则均有好的表现。综合比较表 1 与表 2, 可以发现随着样本量的增加: (1)收敛容差为 10^{-6} 和 10^{-8} 的 dp、ll、以及 Gcomp 方法在 LL_{Best}、LL_{mean}、LL_{max}、LL_{min} 等指标上的表现都在变好, 但是相对而言收敛容差为 10^{-8} 时各方法的表现更好; (2) Itr_{mean}、Itr_{max} 及 λ_{out} 指标均在变小, 甚至 $N = 4000$ 时 λ_{out} 等于 0; (3)在 $N = 500$ 和 1000 水平下, *mCDM* 框架下各收敛准则的表现优于 *GDINA* 框架, 但 $N = 4000$ 时 *GDINA* 框架下大部分收敛准则的表现与 *mCDM* 框架基本一致。

表 3 中呈现的是数据生成模型为饱和模型, $J = 32$ 水平下的模拟研究结果, 由于所有重复中没有极端值, 因此表中没有呈现 λ_{out} 列。另外, 通过表 1 及表 2 中的结果可以发现, 各收敛判断方法在 10^{-8} 的收敛容差水平下的表现明显优于 10^{-6} 水平, 因此不再呈现 dp、ll 类方法在 10^{-6} 的收敛容差水平下的全部结果, 仅保留 mcomp 方法下的结果用于比较说明。表 3 同样显示在 LL_{Best} 指标上表现最好的是 mcomp8。也就是说, 模型完全正确设定条件下, 在本文所探讨的全部收敛准则中, mcomp8 表现最好。

对比表 3 中不同样本量条件下各个收敛准则的表现, 可以发现: (1) $N = 500$ 时, *mCDM* 框架下的大多数收敛判断方法在 LL_{Best} 指标上的表现优于 *GDINA* 框架; $N = 1000$ 和 4000 时, *GDINA* 框架下各收敛准则的表现在变好, 且当收敛容差为 10^{-8} 时 *mCDM* 和 *GDINA* 框架下各收敛准则均有好的表现。(2)就 Itr_{mean}、Itr_{max} 指标而言, 随着样本量的增大, 这两个指标在变小。这说明 $J = 32$ 时, 随着样本量的增大, 所需要的迭代次数在变少。

表 2　饱和 CDM 生成数据, $J = 16$, $N = 1000$ 及 4000 条件下的模拟结果

N	收敛准则	LL_{Best}	LL_{mean}	LL_{max}	LL_{min}	LL_{sd}	t_{mean}	Itr_{mean}	Itr_{max}	λ_{out}
1000	Gdp6	457	−9929.201	−9801.836	−10105.797	49.742	1.057	291	1924	6
	Gdp8	487	−9929.201	−9801.836	−10105.797	49.742	1.660	508	6609	6
	Gll6	117	−9929.201	−9801.836	−10105.797	49.742	0.831	217	713	6
	Gll8	481	−9929.201	−9801.836	−10105.797	49.742	1.107	324	2512	6
	Gcomp6	457	−9929.201	−9801.836	−10105.797	49.742	1.066	295	1924	6
	Gcomp8	487	−9929.201	−9801.836	−10105.797	49.742	1.666	511	6609	6
	mdp6	460	−9929.201	−9801.836	−10105.797	49.742	0.468	288	1950	6
	mdp8	499	−9929.201	−9801.836	−10105.797	49.742	0.726	503	6628	6
	mll6	104	−9929.201	−9801.836	−10105.797	49.742	0.362	213	795	6
	mll8	494	−9929.201	−9801.836	−10105.797	49.742	0.489	323	2509	6
	mcomp6	461	−9929.201	−9801.836	−10105.797	49.742	0.471	291	1950	6
	mcomp8	500	−9929.201	−9801.836	−10105.797	49.742	0.728	507	6628	6
4000	Gdp6	469	−39831.617	−39539.020	−40187.183	102.360	2.588	223	321	0
	Gdp8	500	−39831.617	−39539.020	−40187.183	102.360	3.840	354	506	0
	Gll6	200	−39831.617	−39539.020	−40187.183	102.360	2.334	195	282	0
	Gll8	499	−39831.617	−39539.020	−40187.183	102.360	2.947	261	376	0
	Gcomp6	475	−39831.617	−39539.020	−40187.183	102.360	2.596	224	322	0
	Gcomp8	500	−39831.617	−39539.020	−40187.183	102.360	3.825	356	511	0
	mdp6	463	−39831.617	−39539.020	−40187.183	102.360	1.612	209	312	0
	mdp8	500	−39831.617	−39539.020	−40187.183	102.360	2.376	341	490	0
	mll6	177	−39831.617	−39539.020	−40187.183	102.360	1.443	182	257	0
	mll8	499	−39831.617	−39539.020	−40187.183	102.360	1.774	247	352	0
	mcomp6	471	−39831.617	−39539.020	−40187.183	102.360	1.619	211	312	0
	mcomp8	500	−39831.617	−39539.020	−40187.183	102.360	2.372	342	490	0

表 3　饱和 CDM 生成数据, $J = 32$ 条件下的模拟结果

N	收敛准则	LL_{Best}	LL_{mean}	LL_{max}	LL_{min}	LL_{sd}	t_{mean}	Itr_{mean}	Itr_{max}
500	Gdp8	485	−9334.716	−9163.342	−9521.124	61.640	0.551	77	311
	Gll8	484	−9334.716	−9163.342	−9521.124	61.640	0.452	53	328
	Gcomp8	485	−9334.716	−9163.342	−9521.124	61.640	0.552	77	328
	mdp8	500	−9334.716	−9163.342	−9521.124	61.640	0.235	77	619
	mll8	499	−9334.716	−9163.342	−9521.124	61.640	0.203	54	609
	mcomp6	492	−9334.716	−9163.342	−9521.124	61.640	0.205	52	320
	mcomp8	500	−9334.716	−9163.342	−9521.124	61.640	0.235	77	619
1000	Gdp8	500	−18731.384	−18516.735	−19016.929	93.430	0.682	65	95
	Gll8	500	−18731.384	−18516.735	−19016.929	93.430	0.574	47	66
	Gcomp8	500	−18731.384	−18516.735	−19016.929	93.430	0.682	65	95
	mdp8	500	−18731.384	−18516.735	−19016.929	93.430	0.315	64	95
	mll8	500	−18731.384	−18516.735	−19016.929	93.430	0.266	46	66
	mcomp6	498	−18731.384	−18516.735	−19016.929	93.430	0.263	44	64
	mcomp8	500	−18731.384	−18516.735	−19016.929	93.430	0.315	64	95
4000	Gdp8	500	−75137.975	−74638.007	−75645.526	185.720	1.998	60	71
	Gll8	500	−75137.975	−74638.007	−75645.526	185.720	1.811	48	55
	Gcomp8	500	−75137.975	−74638.007	−75645.526	185.720	1.993	60	71
	mdp8	500	−75137.975	−74638.007	−75645.526	185.720	1.463	58	72
	mll8	500	−75137.975	−74638.007	−75645.526	185.720	1.210	46	56
	mcomp6	489	−75137.975	−74638.007	−75645.526	185.720	1.108	39	50
	mcomp8	500	−75137.975	−74638.007	−75645.526	185.720	1.457	58	72

4.4.2 HCDM 生成数据时各收敛准则的表现

表 4 到表 6 呈现的是通过 HCDM (前 3 个属性是线性层级关系)生成作答反应数据但使用饱和 CDM 估计模型参数条件下的模拟结果。

根据表 4 中的结果，可以发现在所有收敛准则中，*mCDM* 框架下 mcomp8 的表现是最好的，相同收敛容差下 mdp 的表现接近 mcomp。根据 LL_{Best}、LL_{mean}、LL_{max}、LL_{min} 指标，可以发现 *mCDM* 框架下各收敛准则的表现远优于 *GDINA* 框架。就收敛容差而言，*mCDM* 框架下收敛容差的 3 个水平下，各个方法的 LL_{Best}、LL_{mean}、LL_{max}、LL_{min} 等指标表现有明显差异。随着收敛容差的变小，各个方法的表现也在变好，10^{-8} 下各个方法的表现是最佳的。结合表 1，可以发现表 4 中各收敛准则在收敛容差的 10^{-6} 与 10^{-8} 两个水平下的表现的差异更加明显。就迭代次数而言，*mCDM* 及 *GDINA* 下表现最好的 comp8 的最大迭代次数均大于 10000 次，这说明在样本量较小的条件下(如 $N = 500$)，如果将迭代次数设置的过小(如，小于 10000)模型参数估计程序可能会输出不收敛的错误结论。由表 4 中的 λ_{out} 指标可知，在 500 次循环中 Gcomp8 中有 591 个参数存在极端值问题，*mCDM* 框架下的极端值数量为 483。这说明，尽管 *mCDM* 框架能有效减少极端值数量，但是与表 1 中的极端值数量进行对比可以发现边界值问题对于 2 种框架下的模型参数均产生较为负面的影响。

综合表 1 与表 4，在模型中存在边界值条件下同样发现：(1)dp 与 ip 方法的表现具有较高的一致性，且 dp 的表现与 ip 相当或优于 ip；(2)收敛容差等于 10^{-4} 时各收敛判断方法的表现，均没有优于 10^{-6} 或是 10^{-8} 这两个收敛容差下的表现。因此，接下来不再呈现模型中存在边界值条件下 ip 方法及收敛容差等于 10^{-4} 时各收敛准则的结果。

由表 5 中 $N = 1000$ 及 4000 水平下的模拟结果可知，在 LL_{Best} 指标上表现最佳仍然是 mcomp8，其次是 mdp8。就收敛容差而言，各个收敛准则在

表 4 HCDM 生成数据, $J = 16$, $N = 500$ 条件下的模拟结果

收敛准则	LL_{Best}	LL_{mean}	LL_{max}	LL_{min}	LL_{sd}	t_{mean}	Itr_{mean}	Itr_{max}	λ_{out}
Gdp4	1	−4775.050	−4640.212	−4885.902	39.080	0.560	184	870	585
Gdp6	22	−4775.034	−4640.210	−4885.901	39.076	1.276	500	5131	589
Gdp8	27	−4775.033	−4640.210	−4885.901	39.075	2.175	937	23818	591
Gip4	1	−4775.051	−4640.212	−4885.904	39.081	0.543	176	795	585
Gip6	21	−4775.034	−4640.210	−4885.901	39.075	1.231	485	5141	589
Gip8	27	−4775.033	−4640.210	−4885.901	39.075	2.110	922	23818	591
Gll4	0	−4775.048	−4640.214	−4885.902	39.080	0.516	161	714	584
Gll6	12	−4775.036	−4640.210	−4885.901	39.074	0.833	308	1461	588
Gll8	25	−4775.033	−4640.210	−4885.901	39.075	1.284	535	6486	589
Gcomp4	1	−4775.048	−4640.212	−4885.902	39.080	0.574	189	870	588
Gcomp6	22	−4775.034	−4640.210	−4885.901	39.076	1.279	501	5141	589
Gcomp8	27	−4775.033	−4640.210	−4885.901	39.075	2.179	939	23818	591
mdp4	4	−4774.975	−4639.179	−4885.899	39.103	0.221	185	739	486
mdp6	350	−4774.968	−4639.178	−4885.898	39.100	0.403	475	4339	484
mdp8	469	−4774.964	−4639.178	−4885.898	39.101	0.686	931	14029	483
mip4	4	−4774.975	−4639.179	−4885.901	39.103	0.214	179	735	490
mip6	343	−4774.968	−4639.178	−4885.898	39.100	0.387	464	4303	484
mip8	469	−4774.964	−4639.178	−4885.898	39.101	0.647	916	14029	483
mll4	0	−4774.980	−4639.184	−4885.898	39.102	0.201	161	910	473
mll6	72	−4774.969	−4639.178	−4885.898	39.100	0.292	312	1471	482
mll8	458	−4774.965	−4639.178	−4885.898	39.101	0.431	558	5066	486
mcomp4	4	−4774.974	−4639.179	−4885.898	39.103	0.223	191	910	481
mcomp6	351	−4774.968	−4639.178	−4885.898	39.100	0.404	479	4339	484
mcomp8	473	−4774.964	−4639.178	−4885.898	39.101	0.684	936	14029	483

表 5　HCDM 生成数据, $J = 16$, $N = 1000$ 及 4000 条件下的模拟结果

N	收敛准则	LL_{Best}	LL_{mean}	LL_{max}	LL_{min}	LL_{sd}	t_{mean}	Itr_{mean}	Itr_{max}	λ_{out}
1000	Gdp6	9	−9577.383	−9408.520	−9787.279	56.515	1.547	450	5095	491
	Gdp8	12	−9577.379	−9408.520	−9787.279	56.515	2.667	843	17947	494
	Gll6	3	−9577.389	−9408.520	−9787.279	56.510	1.054	285	1685	491
	Gll8	11	−9577.385	−9408.520	−9787.279	56.509	1.558	476	5786	495
	Gcomp6	9	−9577.383	−9408.520	−9787.279	56.515	1.546	451	5095	491
	Gcomp8	12	−9577.379	−9408.520	−9787.279	56.515	2.672	844	17947	494
	mdp6	366	−9577.314	−9408.518	−9787.279	56.508	0.635	467	5512	416
	mdp8	484	−9577.313	−9408.518	−9787.279	56.508	1.171	969	18411	411
	mll6	78	−9577.319	−9408.518	−9787.279	56.503	0.410	285	1686	409
	mll8	470	−9577.319	−9408.518	−9787.279	56.503	0.647	510	5843	415
	mcomp6	370	−9577.314	−9408.518	−9787.279	56.508	0.636	469	5512	416
	mcomp8	488	−9577.313	−9408.518	−9787.279	56.508	1.173	972	18411	411
4000	Gdp6	14	−38423.227	−38076.036	−38778.783	117.696	6.011	604	3920	424
	Gdp8	23	−38423.225	−38076.036	−38778.783	117.696	10.439	1132	12509	427
	Gll6	5	−38423.228	−38076.036	−38778.783	117.696	3.937	375	2066	425
	Gll8	22	−38423.226	−38076.036	−38778.783	117.697	6.492	698	4557	425
	Gcomp6	14	−38423.227	−38076.036	−38778.783	117.696	6.082	612	3920	425
	Gcomp8	23	−38423.225	−38076.036	−38778.783	117.696	10.473	1141	12509	427
	mdp6	276	−38423.146	−38076.034	−38778.782	117.698	3.437	595	3957	356
	mdp8	473	−38423.145	−38076.034	−38778.782	117.698	6.393	1233	12714	355
	mll6	28	−38423.146	−38076.034	−38778.782	117.697	2.253	374	2076	357
	mll8	460	−38423.145	−38076.034	−38778.782	117.698	3.831	733	4569	355
	mcomp6	276	−38423.146	−38076.034	−38778.782	117.698	3.472	602	3957	356
	mcomp8	478	−38423.145	−38076.034	−38778.782	117.698	6.424	1241	12714	355

10^{-6} 与 10^{-8} 水平下的 LL_{Best} 和 LL_{mean} 指标上表现出了明显的差异，收敛容差为 10^{-8} 时这两个指标更好。模型参数估计框架对各收敛准则的表现产生了明显影响，整体而言，在本研究使用的所有指标上，*mCDM* 框架下各收敛准则的表现优于 *GDINA* 框架。样本量对 $J = 16$ 且模型中存在边界值时的各收敛准则在 λ_{out} 指标上的表现同样产生了影响，综合表 4 与表 5，可以发现随着样本量的增加各收敛准则对应的 λ_{out} 的数量在下降。就 Itr_{max} 而言，模型中存在边界值时，*mCDM* 和 *GDINA* 框架下表现较好的模型收敛准则中需要的迭代次数都非常大。例如，$N = 4000$ 条件下 mcomp8 需要的最大迭代次数为 12714，Gcomp8 需要的迭代次数是 12509，这远超 *CDM* 或 *GDINA* 软件包中默认的迭代次数。

通过表 4 与表 5 中的结果可知，模型估计框架为 *mCDM*、收敛容差为 10^{-8} 时各收敛准则的表现更好。因此，表 6 中不再呈现 *GDINA* 框架及 10^{-4}、10^{-6} 收敛容差下完整的模拟结果，仅包含 Gcomp8 及 mcomp6 用于结果比较。表 6 中呈现的是模型中存在边界值且 $J = 32$ 时的模拟结果。可以发现在 $N = 500$、1000 及 4000 这 3 个样本量水平下表现最好的收敛准则都是 mcomp8, mdp8 与 mll8 的表现相对较好。就模型参数估计框架而言：(1) *mCDM* 框架下各收敛准则在 LL_{Best}、LL_{mean}、t_{mean}、λ_{out} 等指标上的表现优于相同收敛准则在 *GDINA* 框架下的表现；(2) *mCDM* 框架下收敛容差的值对各收敛方法的表现有明显影响，收敛容差的值越小，同一收敛判断方法在 LL_{Best}、LL_{mean} 指标上的表现越好。样本量对于各收敛准则对应的 λ_{out} 指标产生了明显的影响，同一种收敛准则下样本量越大 λ_{out} 的值越小。对比相同收敛准则在 $J = 16$ (表 4 与表 5) 及 $J = 32$ 水平下的 Itr_{mean} 和 Itr_{max} 指标上的表现，可以发现随着项目量增大 Itr_{mean} 和 Itr_{max} 在下降；然而需要特别指出的是，即使在 $J = 32$ 条件下 Gcomp8 和 mcomp8 收敛准则中的 Itr_{max} 仍可能大于 3000。

表 6 HCDM 生成数据, $J = 32$ 条件下的模拟结果

N	收敛准则	LL_{Best}	LL_{mean}	LL_{max}	LL_{min}	LL_{sd}	t_{mean}	Itr_{mean}	Itr_{max}	λ_{out}
500	Gcomp8	83	−8944.542	−8746.172	−9101.048	63.686	0.823	143	4521	1072
	mdp8	416	−8944.529	−8746.349	−9100.836	63.714	0.309	162	3678	936
	mll8	417	−8944.529	−8746.349	−9100.836	63.714	0.241	109	1701	916
	mcomp6	390	−8944.531	−8746.349	−9100.836	63.713	0.240	101	1575	921
	mcomp8	417	−8944.529	−8746.349	−9100.836	63.714	0.310	163	3678	936
1000	Gcomp8	44	−17941.473	−17692.040	−18203.752	96.770	1.375	179	6530	998
	mdp8	456	−17941.322	−17692.038	−18205.384	96.780	0.607	218	12877	810
	mll8	452	−17941.322	−17692.038	−18205.384	96.780	0.411	124	1840	805
	mcomp6	408	−17941.322	−17692.038	−18205.384	96.780	0.420	115	3035	809
	mcomp8	456	−17941.322	−17692.038	−18205.384	96.780	0.610	219	12877	810
4000	Gcomp8	51	−71973.595	−71443.652	−72679.347	198.161	7.854	278	7908	913
	mdp8	443	−71973.490	−71443.649	−72679.344	198.184	5.795	299	6037	714
	mll8	443	−71973.494	−71443.649	−72679.344	198.185	3.729	191	1799	706
	mcomp6	373	−71973.496	−71443.649	−72679.344	198.184	3.470	164	1833	717
	mcomp8	449	−71973.490	−71443.649	−72679.344	198.184	5.896	303	6037	714

5 实证数据分析

数据来源于 Yuan 等人(2022)关于小学数学分数运算的认知诊断研究。这个数据集包含 817 名被试对 56 个项目的作答。Yuan 等人(2022)在文献分析的基础上，根据专家建议、被试访谈及口语报告法等，定义了 5 个认知属性，分别是：基本运算(α_1)、约分(α_2)、通分(α_3)、带分数拆分(α_4)、借位(α_5)。其研究提出分数运算认知过程的可能路径是：掌握 α_1 是掌握 α_2、α_3、α_5 的前提；由于属性 α_4 仅涉及将整数与分数部分拆开，不需要预先掌握 α_1；图 4 中呈现了认知属性层级关系图。Yuan 等人(2022)使用似然比统计量比较了 logit 连接函数下饱和 CDM 与 HCDM 的对数似然函数值的差异，初步证实了小学数学分数运算数据集中存在图 4 中所呈现的层级关系。

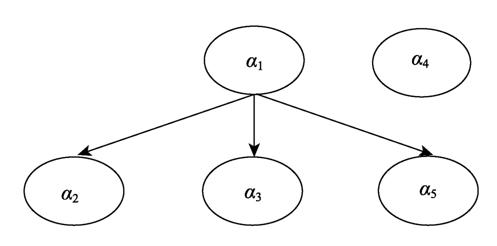


图 4 Yuan 等人(2022)定义的小学数学分数运算认知属性层级关系

本文以小学数学分数运算数据集为例，探讨当 CDM 模型中存在边界值时，*GDINA* 及 *mCDM* 模型参数估计框架下，5 种收敛判断方法(dp、ip、ll、rl、comp)，3 种收敛容差(10^{-4}、10^{-6}、10^{-8})，所组成的 30 种收敛判断准则的表现。表 7 中呈现了这 30 种收敛准则对应的对数似然函数值(简记为，LL)，以秒为单位的模型参数估计时间(t)，迭代次数以及 λ_{out}；为便于结果解释，将 LL 值保留到了小数点后四位。

根据模型参数估计的极大似然理论，收敛判断准则对应的 LL 越大，说明这个准则的表现越好，模型参数点估计值的可靠性越高。

可以发现：(1)对于 LL 值影响最大的是模型参数估计框架，本研究中新开发的 *mCDM* 框架下各收敛准则对应的 LL 值远大于 *GDINA* 框架下各收敛准则对应的 LL 值。(2)在所有收敛准则中表现最好的是 mdp8 与 mcomp8，在这两种收敛准则中不仅似然函数是最大的，而且项目参数中没有极端值。(3)就 3 种收敛容差而言，不论是在 *mCDM* 还是 *GDINA* 框架下，10^{-4} 的表现都是最差的，10^{-8} 的表现是最佳的；尽管在一些收敛准则中 10^{-6} 的表现与 10^{-8} 类似，但是前者并不具有普遍适用性。以上 3 个发现与模拟研究中的结论具有高度的一致性。

6 讨论与展望

本文通过理论分析及模拟研究证实，心理计量模型的点估计值在一些情景中会存在可靠性问题，且新开发的模型参数估计框架及收敛准则能够提高模型参数估计值的可靠性。

表 7　实证数据分析结果

GDINA 框架					mCDM 框架				
收敛准则	LL	t	Itr	λ_{out}	Cov	LL	t	Itr	λ_{out}
Gdp4	−14307.9718	1.040	133	4	mdp4	−14248.5465	0.470	64	1
Gdp6	−14307.9717	1.328	190	4	mdp6	−14248.5463	0.718	111	1
Gdp8	−14307.9717	1.686	247	4	mdp8	−14248.5463	0.975	158	0
Gip4	−14307.9719	0.914	123	4	mip4	−14248.5469	0.423	58	0
Gip6	−14307.9717	1.299	181	4	mip6	−14248.5463	0.670	105	1
Gip8	−14307.9717	1.631	238	4	mip8	−14248.5463	0.925	152	1
Gll4	−14307.9720	0.891	119	4	mll4	−14248.5465	0.449	63	3
Gll6	−14307.9717	1.128	148	4	mll6	−14248.5463	0.570	87	1
Gll8	−14307.9717	1.245	177	4	mll8	−14248.5463	0.698	110	2
Grl4	−14351.6261	0.264	20	4	mrl4	−14257.7213	0.168	13	0
Grl6	−14308.0450	0.448	47	4	mrl6	−14248.6033	0.289	35	1
Grl8	−14307.9725	0.856	111	4	mrl8	−14248.5469	0.415	58	0
Gcomp4	−14307.9718	1.040	133	4	mcomp4	−14248.5465	0.470	64	1
Gcomp6	−14307.9717	1.328	190	4	mcomp6	−14248.5463	0.718	111	1
Gcomp8	−14307.9717	1.686	247	4	mcomp8	−14248.5463	0.975	158	0

6.1　讨论

首先，通过预研究作者认为最大迭代次数设置过少可能会导致模型参数不收敛的问题(如，3000或以下，见 *GDINA* 及 *CDM* 软件包)，因此本研究将最大迭代次数设置为 50000。模拟研究发现，本文所有实验条件组合下 *mCDM* 和 *GDINA* 这两种模型参数估计框架均收敛。模拟研究显示在一些特定条件下(见表 1)，*mCDM* 和 *GDINA* 的最大迭代次数均超过了 30000 次，这也就意味着如果将最大收敛次数设置为 3000 那么就会出现模型参数不收敛的问题。因此，本文认为增大模型参数估计程序的最大迭代次数有助于解决模型参数不收敛问题。

其次，针对 CDM 中可能存在的边界值以及项目参数存在极端值问题，本文开发了新的 CDM 模型参数估计框架 *mCDM*。通过对比 *mCDM* 和 *GDINA* 这两种模型参数估计框架在模拟研究及实证数据分析中的表现，发现 *mCDM* 框架的表现优于或至少与 *GDINA* 框架的表现相当；且 *mCDM* 框架有效减少了项目参数极端值数量。因此，本文认为在估计 CDM 模型参数时，*mCDM* 可能是一个更好的选择。导致 CDM 中存在边界值的一个原因是属性间存在层级关系，使得饱和 CDM 中的一些参数近似等于 0。研究者以饱和 CDM 为基础开发了一些属性层级关系探索或验证的方法(Gu & Xu 2019; Liu et al., 2022; Templin & Bradshaw, 2014)。我们建议研究者进一步在 *mCDM* 框架下使用已有方法或者是开发新方法对属性层级关系进行研究。当有较为充分的证据证明层级关系存在时，在 *mCDM* 框架下使用 HCDM 分析数据，可能会提高模型参数点估计值的可靠性。

第三，本文新提出模型参数收敛综合判断法 comp，并在 2 种参数估计框架(*mCDM* 和 *GDINA*)、3 种收敛容差(10^{-4}、10^{-6}、10^{-8})下比较了 dp、ip、ll、rl 及 comp 等方法所组成的 30 种收敛准则的表现。就本研究所探讨的 3 种收敛容差而言，10^{-8} 的表现是最好的，10^{-4} 的表现则不及 10^{-6} 和 10^{-8}；收敛容差的值越小收敛准则的表现越好，尤其是在 *mCDM* 框架下。就 dp、ip、ll、rl 及 comp 这 5 种收敛判断方法而言，comp 的表现最好，rl 方法的表现最差；在 *mCDM* 框架下表现最为明显。因此，本文认为，估计模型参数时，*mCDM* 框架下收敛容差为 10^{-8} 的 comp 方法的可靠性较高。

6.2　展望

本文以同一连接下的饱和 G-DINA 模型为例，探讨了 *mCDM* 和 *GDINA* 框架下目前已有的及本研究新开发的各收敛准则在 CDM 模型参数估计中的表现。尽管本研究初步解决了在 CDM 模型参数估计时如何选择恰当收敛准则的问题，但是作者认为有以下几个问题需要进一步探索。

第一个是关于 rl 方法所适用的收敛容差问题。本文发现，相对于其他准则而言，rl 取 10^{-4}、10^{-6}、10^{-8} 这 3 种收敛容差值时在 LL_{Best}、LL_{mean} 指标上

的表现均较差。通过 rl 的计算方式 $\text{abs}\{[\ell(\gamma^{(rep+1)}|y)-\ell(\gamma^{(rep)}|y)]/\ell(\gamma^{(rep+1)}|y)\}$，结合 ll 方法可以分析出这个问题出现的原因。以表 1 中呈现的模拟结果为例，可以发现在这个实验条件组合下，ll 方法在 *mCDM* 框架下且收敛准则值等于10^{-8}时的表现，相对于其他框架及收敛容差较优。根据 mll8 的定义，此时 $\text{abs}\{-2[\ell(\gamma^{(rep+1)}|y)-\ell(\gamma^{(rep)}|y)]\}<10^{-8}$；再根据表中的 LL_{mean} 值−4948.008，可近似获得此时 mll 的值，并计算 mrl，

$$\text{mrl}=\text{abs}\{[\ell(\gamma^{(rep+1)}|y)-\ell(\gamma^{(rep)}|y)]/\ell(\gamma^{(rep+1)}|y)\}<10^{-8}/(2\times4948.008) \quad (2)$$

这也就意味着，如果 mrl 想要达到与 mll8 相近的效果，mrl 方法的收敛容差应该近似等于10^{-12}。因此，作者建议后续研究者可以沿着这个线索继续探索 rl 方法的表现。

第二是关于 *mCDM* 框架及其应用的问题。本研究开发 *mCDM* 框架的主要目的在于提供一个更加合理的 CDM 模型参数估计框架，尽量减少模型参数不收敛、边界值问题及项目参数极端值对 CDM 模型参数收敛准则表现的影响。特别说明的是模拟实验中将最大迭代次数设置为 50000 时，两种参数估计框架下的所有循环中的参数估计都收敛了，因此在本研究中 *mCDM* 框架仅在边界值问题及项目参数存在极端值时起作用。模型中存在边界值时，尽管 *mCDM* 框架下的项目极端值数量少于同条件下 *GDINA* 框架所对应的数量，但即使是在 $N=4000$ 条件下，*mCDM* 框架下出现极端值的频率仍然较高。因此，本研究认为有必要以 *mCDM* 框架为基础，继续对模型参数不收敛、边界值问题及项目参数极端值等问题展开探索。

第三，不同连接函数下各种收敛准则的表现有待进一步探索。本文以同一连接下的饱和 G-DINA 模型为例，探讨了不同收敛准则的表现。但 CDM 中还有两种得到广泛应用的连接：logit 连接以及 log 连接(de la Torre, 2009, 2011; Templin & Bradshaw, 2014)。这 3 种连接函数的主要区别之一是，项目参数与项目正确作答概率之间关系的表达不同。鉴于 dp 的表现在大多数情况下略优于 ip，本研究认为后续研究可以对不同连接函数下各个收敛准则的表现展开进一步探索。

参 考 文 献

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452−454.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483* (7391), 531−533.

Chiu, C. Y., Köhn, H. F., & Ma, W. (2023). Commentary on "Extending the Basic Local Independence Model to Polytomous Data" by Stefanutti, de Chiusole, Anselmi, and Spoto. *Psychometrika, 88*(2), 656−671.

DeCarlo, T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-Matrix. *Applied Psychological Measurement, 35*(1), 8−26.

DeCarlo, T. (2019). Insights from reparameterized DINA and beyond. In M. von Davier & Y.-S. Lee (Eds.). *Handbook of diagnostic classification models* (pp. 549−572). Springer.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115−130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179−199.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1−22.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1−24.

Gu, Y., & Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research, 20*(115), 1−58.

Gu, Y., & Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics, 48*(4), 2082−2107.

Hu, C., Wang, F., Guo, J., Song, M., Sui, J., & Peng. K. (2014). The replication crisis in psychological research. *Advances in Psychological Science*, 24(9), 1504−1518.

[胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题: 从危机到契机. *心理科学进展, 24*(9), 1504−1518.]

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640−648.

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software *mdltm* Including parallel EM algorithm. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 603−628). Springer.

Liu, R. (2018). Misspecification of attribute structure in diagnostic measurement. *Educational and psychological measurement*, *78*(4), 605−634.

Liu, Y. (2022). Standard errors and confidence intervals for cognitive diagnostic models: Parallel bootstrap methods. *Acta Psychologica Sinica, 54*(6), 703−724.

[刘彦楼. (2022). 认知诊断模型的标准误与置信区间估计: 并行自助法. *心理学报, 54*(6), 703−724.]

Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*(1), 3−26.

Liu, Y., Xin, T., & Jiang, Y. (2022). Structural parameter

standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research*, *57*(5), 784−803.
Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253−275.
Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*(14), 1−26.
Ma, W., de la Torre, J., Sorrel, M., & Jiang, Z. (2022). *GDINA: The generalized DINA model framework*. R package version 2.9.3. https://CRAN.R-project.org/package=GDINA
Ma, W., & Guo, W. (2019). Cognitive diagnosis models for multiple strategies. *British Journal of Mathematical and Statistical Psychology, 72*(2), 370−392.
Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied Psychological Measurement*, *45*(2), 95−111.
Paek, I., & Cai, L. (2013). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement, 74*(1), 58−76.
Paulsen, J., & Valdivia, D. S. (2022). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education*, *90*(4), 916−933.
Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, *43*(1), 88−115.
Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2022). *CDM: Cognitive Diagnosis Modeling*. R package version 8.2-6. http://CRAN.R-project.org/package=CDM
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
Rupp, A. A., & van Rijn, P. W. (2018). GDINA and CDM packages in R. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 71−77.
Sen, S., & Terzi, R. (2020). A comparison of software packages available for dina model estimation. *Applied Psychological Measurement*, *44*(2), 150−164.
Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgment test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*(3), 506−532.
Tajika, A., Ogawa, Y., Takeshima, N., Hayasaka, Y., & Furukawa, T. A. (2015). Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *The British Journal of Psychiatry, 207*(4), 357−362.
Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317−339.
Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using *Mplus*. *Educational Measurement: Issues and Practice, 32*(2), 37−50.
Tian, W., Xin, T., & Kang, C. (2014). The data-augmentation techniques in item response modeling: Current approaches and new developments. *Advances in Psychological Science*, *22*(6), 1036−1046.
[田伟, 辛涛, 康春花. (2014). 项目反应理论中潜在心理特质“填补”的参数估计方法及其演变. *心理科学进展*, *22*(6), 1036−1046.]
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287−307.
Wu, Z., Deloria-Knoll, M., & Zeger, S. L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics, 18*(2), 200−213.
Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data*. *ETS Research Report Series*, *2008*(1), i−18.
Yamaguchi, K. (2023). On the boundary problems in diagnostic classification models. *Behaviormetrika, 50*(1), 399−429.
Yuan, L., Liu, Y., Chen, P., & Xin, T. (2022). Development of a new learning progression verification method based on the hierarchical diagnostic classification model: Taking grade 5 students' fractional operations as an example. *Educational Measurement: Issues and Practice, 41*(3), 69−82.
Zeng, Z., Gu, Y., & Xu, G. (2023). A Tensor-EM method for large-scale latent class analysis with binary responses. *Psychometrika, 88*(2), 580−612.

On the reliability of point estimation of model parameters: Taking cognitive diagnostic models as an example

LIU Yanlou[1,2], CHEN Qishan[3,4], WANG Yiming[2], JIANG Xiaotong[2]

([1]*Academy of Big Data for Education*; [2]*School of Psychology, Qufu Normal University, Jining* 273165, *China*)

([3]*Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents* (*South China Normal University*), *Ministry of Education;* [4]*School of Psychology, South China Normal University, Guangzhou* 510631, *China*)

Abstract

Cognitive diagnostic models (CDMs) are psychometric models that have received increasing attention within fields such as psychology, education, sociology, and biology. It has been argued that an inappropriate convergence criterion for a maximum likelihood estimation using the expectation maximization (MLE-EM)

algorithm could result in unpredictable and inaccurate model parameter estimates. Thus, inappropriate convergence criteria may yield unstable and misleading conclusions from the fitted CDMs. Although several convergence criteria have been developed, it remains an unexplored question, how to specify the appropriate convergence criterion for fitted CDMs.

A comprehensive method for assessing convergence is proposed in this study. To minimize the influence of the model parameter estimation framework, a new framework adopting the multiple starting values strategy (*mCDM*) is introduced. To examine the performance of the convergence criterion for MLE-EM in CDMs, a simulation study under various conditions was conducted. Five convergence assessment methods were examined: the maximum absolute change in model parameters, the maximum absolute change in item endorsement probabilities and structural parameters, the absolute change in log-likelihood, the relative log-likelihood, and the comprehensive method. The data generating models were the saturated CDM and the hierarchical CDM. The number of items was set to J = 16 and 32. Three levels of sample sizes were considered: 500, 1000, and 4000. The three convergence tolerance value conditions were 10^{-4}, 10^{-6}, and 10^{-8}. The simulated response data were fitted by the saturated CDM using the *mCDM* and the R package *GDINA*. The maximum number of iterations was set to 50000.

The simulation results suggest the following.

(1) The saturated CDM converged under all conditions. However, the actual number of iterations exceeded 30000 under some conditions, implying that when the predefined maximum iteration number is less than 30000, the MLE-EM algorithm might inadvertently stop.

(2) The model parameter estimation framework affected the performance of the convergence criteria. The performance of the convergence criteria under the *mCDM* framework was comparable or superior to that of the *GDINA* framework.

(3) Regarding the convergence tolerance values considered in this study, 10^{-8} consistently had the best performance in providing the maximum value of the log-likelihood and 10^{-4} had the worst performance. Compared to all other convergence assessment methods, the comprehensive method in general had the best performance, especially under the *mCDM* framework. The performance of the maximum absolute change in model parameters was similar to the comprehensive method, but this good performance was not consistent. On the contrary, the relative log-likelihood had the worst performance under the *mCDM* and *GDINA* frameworks.

The simulation results showed that the most appropriate convergence criterion for MLE-EM in CDMs was the comprehensive method with tolerance 10^{-8} under the *mCDM* framework. The results from the real data analysis also demonstrated that the proposed comprehensive method and *mCDM* framework had good performance.

Keywords model parameter estimation, point estimation, convergence criterion, cognitive diagnostic model